Biodiversity Informatics, 6, 2009, pp. 53-58

EFFECTIVELY SEARCHING SPECIMEN AND OBSERVATION DATA WITH *TOQE*, THE THESAURUS OPTIMIZED QUERY EXPANDER

A. GÜNTSCH, N. HOFFMANN, P. KELBERT AND W. BERENDSOHN
Freie Universität Berlin, Botanic Garden and Botanical Museum Berlin-Dahlem, Königin-Luise-Str. 6-8, D-14195 Berlin, Germany

Abstract.— Today's specimen and observation data portals lack a flexible search mechanism, able to link up thesaurus-enabled data sources such as taxonomic checklist databases and expand user queries to related terms, significantly enhancing result sets. The TOQE system (Thesaurus Optimized Query Expander) is a REST-like XML web-service implemented in Python and designed for this purpose. Acting as an interface between portals and thesauri, TOQE allows the implementation of specialized portal systems with a set of thesauri supporting its specific focus. It is both easy to use for portal programmers and easy to configure for thesaurus database holders who want to expose their system as a service for query expansions. Currently, TOQE is used in four specimen and observation data portals. The documentation is available from http://search.biocase.org/toqe/.

Key words.— BioCASE, GBIF, EDIT, query expansion, specimen portal, SYNTHESYS, taxonomy, thesaurus, XML.

Over the last decade, specimen and observational data portals have been developed with a wide variety of thematic scopes and technologies. However, all share the goal of offering unified access for distributed and often very heterogeneous data sources using standardized data schemas (e.g. ABCD and DarwinCore) or common access protocols (e.g. BioCASE, DiGIR, and TAPIR).

Early systems such as the Species Analyst network (Vieglas, 2003) and the European Natural History Specimen Information Network (ENHSIN, n.d.; Güntsch, 2002) used entirely distributed query mechanisms, instantly propagating user requests to the respective networks. The rapidly growing number of available data providers and data records led to serious performance and accessibility issues, and made this strategy infeasible. Today, the vast majority of specimen and observation portals use index databases containing a projection of the networked concepts, which usually consist of a limited number of elements considered fundamental for typical queries. As a result, core data elements can be processed and retrieved with almost no delay. In a second query step, full data records can be accessed directly from the data provider as needed. The most complete index database, having a worldwide scope and used by several different portal systems, is the GBIF index. Presently it contains 180.000.000 data records (August 2009).

correspondence email: BiodiversityInformatics@bgbm.org.

Apart from the obvious performance and stability benefit, the usage of an index database offers the opportunity to harmonize the highly heterogeneous vocabularies used by distributed and independent data providers. These can include: various naming conventions for scientific organism names; different taxonomic classifications for the same organism; misspelled country codes; varying representations of geographic coordinates; and different spellings of person names (e.g. collectors). These disparities can be partly resolved in the indexing process, for example: by mapping taxonomic names to standard taxonomies; by identifying and correcting potential misspellings; and by translating and converting data formats into standardized and common representations.

However, even after the data stored in a common index database has been harmonized to a large extent, additional knowledge about the potential relations between terms can significantly broaden the result set for a given user query. A taxonomic checklist database might for example contain information about existing misapplications of scientific names. Specimens or observations identified under the misapplied name will not be found if the query term is only the correct name.

There are many specialized checklist or thesaurus databases which could potentially support portals by expanding user queries to related terms, such as taxonomic checklist databases, gazetteers, lists of person names, stratigraphic term lists, etc. It does

not necessarily make sense to use all of them in a comprehensive system such as the GBIF search portal. But specific regional or thematic networks can be improved significantly by using a selected set of thesaurus systems relevant for the given focus. The thesauri should therefore ideally be equipped with a service layer, one which is both easy to implement by the existing thesaurus database and easy to use by portals for query expansions. Portal systems could then choose from the set of available thesauri and invoke them for their specific regional or thematic foci.

The Thesaurus Optimized Query Expander TOQE service was defined and implemented in the context of the EU 6th framework project SYNTHESYS (Synthesis of Taxonomic resources, SYNTHESYS, n.d.), and uses this approach to support queries to the Biological Collection Access Service for Europe (BioCASE, 2005-) with high quality taxonomic data available from Fauna Europaea (Fauna Europaea, 2004), Euro+Med Plantbase (Euro+Med, 2006-), and the European Register of Marine Species (ERMS, 2004). We tried to make the TOQE implementation as generic and configurable as possible with regard to the underlying thesaurus database, which led to many more provider databases and portals than originally planned using the services. In the following, we will describe the TOQE service and its implementation, its primary deployment in the European BioCASE portal, and further implementation beyond the initial scope of the project.

THE TOQE SERVICE

A TOQE service can deliver a set of concepts using a specific term, such as a set of taxa returned by a given scientific name in a floristic database. The service can then be used to retrieve related concepts (e.g. included taxa, misapplied names, synonyms) for the concepts that have been identified in the first step and for the terms being used (fig. 1). The resulting term-list is then used to expand the original user query.

TOQE itself is an XML web-service offering five methods supporting and partly simplifying the two-step query mechanism. A TOQE-method is called with a GET-request. The following example shows a call of the method getConceptsByTerm querying for all concepts in a given thesaurus using the term "Calendula arvensis":

Text box 1:
http://search.biocase.org/toqe/toqe.py?term=Calendula+arv%&thesaurus=Standardliste&method=getConceptsByTerm

The syntax of the corresponding XML response-documents is defined in the TOQE schema[1]. The following response belongs to the above function call:

Text Box 2:

```
<?xml version="1.0" encoding="utf-8" ?>
<response success="true">
   <conceptSet searchTerm="calendula arvensis">
         <concept
         dataSchemaConcept="DataSets/[…]/HigherTaxa/HigherTaxon/HigherTaxonName">
               <name id="10794" relationtype="" id2="" name="">Calendula arvensis @authorL.</name>
               <status>accepted</status>
               <reference id="1">
               R. Wisskirchen ET H. Haeupler - Standardliste der Farn- und Blütenpflanzen Deutschlands 1998.
               </reference>
      </concept>
   </conceptSet>
</response>
```

A full implementation of the TOQE service offers the following five methods:

getMethods() returns the list of methods implemented by a given TOQE implementation.

getMethodInfo(methodName: String) returns a description of a specific TOQE method (required argument methodName).

getConceptsByTerm(term: String, thesaurus: String) receives a search term and the name of a thesaurus connected to the given TOQE instance and returns the list of concepts matching the search term in this thesaurus. The '%' character can be used as a wildcard within the search term. Both arguments are required.

getRelatedConceptsByTerm(term: String, relation: String[], thesaurus: String) is more powerful than getConceptsByTerm and returns both matching concepts (ids, names, references, status values) for the search term and related concepts.

[1] http://search.biocase.org/toqe/schema/

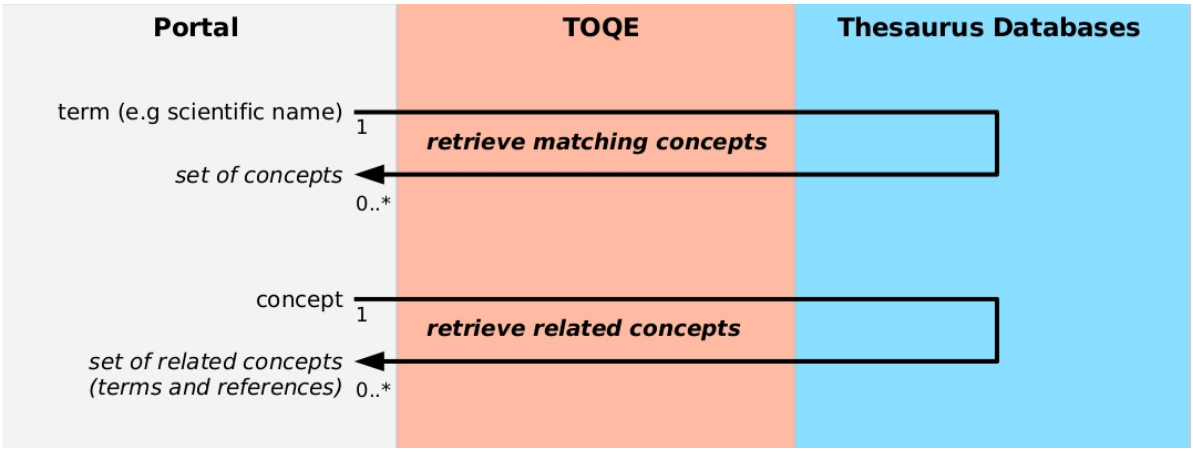

Figure 1. Scheme of how queries are expanded using TOQE.

The relation types to be analyzed can be passed with the repeatable relation argument. Again, the thesaurus to be used must be specified. All arguments are required.

getAllRelatedConcepts(conceptKey: String, relation: String[], recursive: Boolean, thesaurus: String) receives a single concept key, a list of relation types to be analyzed in the given thesaurus and returns all related concepts (ids, names, references, status values). Additionally, the Boolean argument recursive is used to indicate whether the search should return concepts that are not directly related to the given concept and have to be searched recursively. The semantics of a recursive search depends on the thesaurus being queried, as well as on the available computing resources and the capabilities of the database management software in use. An example of a strategy for recursive searches of taxonomic concepts in the Euro+Med Plantbase is here[2]. The recursive argument is optional, all other arguments are required.

A more detailed description of the service, including the set of error-responses, is available here[3]. So far, a registry service for TOQE has not been implemented, so client software systems are required to know the available and appropriate TOQE services.

The TOQE software implemented in the context of the BioCASE portal development has two basic layers (fig. 2). The service layer processes the incoming HTTP GET-request and calls the associated database access method, hiding database specific functionalities in a generic way. The returned records are then wrapped up in XML response documents, following the TOQE schema specification. The native database queries are generated in the TOQE database layer. A database module has to be configured for each thesaurus database connected to the TOQE service layer.

All software components have been implemented using the Python programming language. The source code is freely available and can be obtained from the BGBM subversion repository here[4].

THE BIOCASE PORTAL AND TOQE

In spring 2008, BioCASE released a new data portal which uses the TOQE query expansion service. The BioCASE portal uses a subset of GBIF specimen and observation data records referring to organisms that have been collected or observed in Europe (Holetschek et al. 2006). Accordingly, the thesaurus systems connected with TOQE are the three major European taxonomic checklist systems: Euro+Med plantbase, Fauna Europaea, and European Register of Marine Species (ERMS).

Figure 3 shows the BioCASE response for a user query for Calendula arvensis. The original query is expanded using 2 misapplied names, 32 synonyms,

[2] http://search.biocase.org/bgbm/static/extraSearch/extraSearch.htm.
[3] http://search.biocase.org/toqe/api.html#client.
[4] http://ww2.biocase.org/svn/synthesys/trunk/thesaurus.

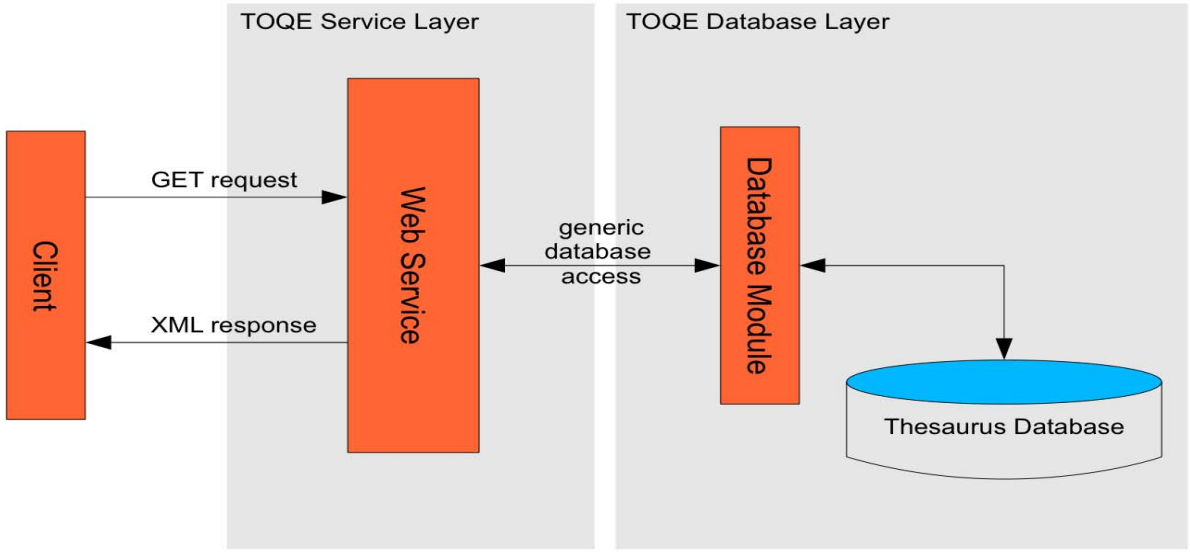

Figure 2. Schematic showing TOQE implementation.

and the higher taxon Calendula,as well as a number of spelling variants. This leads to 3139 hits in the GBIF index, compared to 2844 hits for the original query without query expansion.

The portal returns both the expanded list of terms (scientific names) used in the query and the list of terms with actual hits in the GBIF index database, together with the respective number of units which will be returned. Users can now deselect terms which should not be contained in the specimen result set.

Other Implementations

Because both BioCASE portal software and the TOQE thesaurus service are highly generic and configurable, the implementation of additional portal systems with differing scopes can be accomplished with relatively little effort. Up to now, the following additional TOQE-enabled portals have been developed:

Checklist driven access to European Biodiversity Data (prototype)[5]: Like the standard BioCASE portal, the system gives access to specimens and observations from Europe with expanded queries using the major European floristic and faunistic taxonomic checklist databases. In contrast to the BioCASE portal, users have full control over the query expansion process and can freely select the thesaurus systems to be used and the relation types to be considered.

EDIT Specimen and Observation Explorer for Taxonomists[6]: The portal offers an interface to all GBIF data specifically tailored to serve the taxonomic work process. The TOQE query functions are directly integrated into the specimen search interface so that taxonomists can easily pick from the list of available thesaurus systems. In a second query step, a clearly laid out form summarizes terms and relations retrieved from the selected thesauri for the given query. The system will be an important component of the EDIT Platform for Cybertaxonomy (EDIT, 2007-; Berendsohn et al., 2007; Döring, 2007).

BioCASE Portal for BGBM Collections[7]: The Botanic Garden and Botanical Museum Berlin-Dahlem has linked almost all its collection databases to GBIF, ranging from smaller collection databases for particular herbarium subcollections to the comprehensive accession management system of the Botanical Garden's living collection. This opened up the opportunity to set up a portal specifically dedicated to BGBM collections. So far, the German standard lists for ferns and flowering plants and the reference list for German Bryophytes

[5] http://search.biocase.org/toto/.

[6] http://search.biocase.org/edit/.

[7] http://search.biocase.org/bgbm/.

Figure 3. Query expansion for the term *Calendula arvensis* using TOQE in BioCASE. The original result set is increased from 2844 to 3139 records.

as well as Euro+Med plantbase and ERMS, have been added as thesaurus databases. The new portal offers a convenient new access point to BGBM collections which were previously lacking a common and unified representation.

Conclusions

Expanding queries to specimen and observational databases using the relevant thesaurus systems for a given portal scope can lead to richer query responses, with results which would have been ignored using traditional query mechanisms. We have demonstrated this approach with the definition and implementation of the TOQE-service now in use by several specimen and observation portals together with a variety of taxonomic thesaurus systems. The approach should now be broadened to include non-taxonomic concepts such as place names, adding further value to specimen portals. We also believe that the query expansion process will require a sophisticated interactive explanation component monitoring the query mechanism transparently to the end-user.

Acknowledgements

This work was funded by the Network Activity D of the European Union 6th Framework project SYNTHESYS (contract RII3-CT-2003-506117). The authors thank Jörg Holetschek, Wolf-Henning Kusber and Elke Zippel of the SYNTHESYS project team for their critical comments and assistance during the implementation phase. We would also like to thank the Fauna Europaea, ERMS, and Euro+MED initiatives for the unbureaucratic provision of data. Our special thanks go to Pepe Ciardelli for proofreading of the manuscript.

Literature Cited

Berendsohn, W. G., M. Döring, and M. C. Ebach 2007. EDIT needs Biodiversity Information Standards. P. 1 in: Weitzman, A., and L. Belbin ed. Abstracts of the 2007 annual conference of the Taxonomic Databases Working Group in Bratislava.

BioCASE 2005-. Biological Collection Access Service. Accessed 20 August 2008 from http://www.biocase.org.

Döring, M. 2007. A general concept for the design of the EDIT Platform for Cybertaxonomy. EDIT Newsletter 3. Muséum National d'Histoire Naturelle, Paris.

EDIT 2007-. EDIT platform for cybertaxonomy. Accessed 20 August 2008 from http://wp5.e-taxonomy.eu/blog/index.php.

ENHSIN n.d. European Natural History Specimen Information Network. Accessed 20 August 2008 from http://www.nhm.ac.uk/research-curation/projects/ENHSIN/index.html.

ERMS 2004. The European Register of Marine Species. Accessed 20 August 2008 from http://www.marbef.org/data/erms.php.

Euro+Med 2006-. Euro+Med Plantbase. Accessed 20 August 2008 from http://ww2.bgbm.org/EuroPlusMed/query.asp.

Fauna Europaea 2004. Fauna Europaea. Accessed 20 August 2008 from http://www.faunaeur.org/.

Güntsch, A. 2002. The ENHSIN pilot network. P.p. 33-40 in Scoble, M. J. ed. ENHSIN – The European Natural History Specimen Information Network. The Natural History Museum, London.

Holetschek, J., A. Güntsch, C. Oancea, M. Döring, and W. G. Berendsohn 2006. Prototyping a Generic Slice Generation System for the GBIF Index. Pp. 51-52 in Belbin, L., A. Rissoné, and A. Weitzman, A. eds. Abstracts of the 2006 annual conference of the Taxonomic Databases Working Group in St Louis, MI.

SYNTHESYS n.d. Synthesis of Taxonomic Resources. Accessed 20 August 2008 from http://www.synthesys.info/index.htm.

Vieglas, D. 2003. Species Analyst. Accessed 20 August 2008 from http://www.faunaeur.org/.